Soc Indic Res (2008) 86:481–496
DOI 10.1007/s11205-007-9181-8

On Exact Statistical Properties of Multidimensional Indices Based on Principal Components, Factor Analysis, MIMIC and Structural Equation Models

Jaya Krishnakumar · A. L. Nagar

Accepted: 13 August 2007 / Published online: 4 September 2007

Abstract Recent empirical literature has seen many multidimensional indices emerge as well-being or poverty measures, in particular indices derived from principal components and various latent variable models. Though such indices are being increasingly and widely employed, few studies motivate their use or report the standard errors or confidence intervals associated with these estimators. This paper reviews the different underlying models, reaffirms their appropriateness in this context, examines the statistical properties of resulting indices, gives analytical expressions of their variances and establishes certain exact relationships among them.

Keywords Principal components · Factor analysis · Structural equation models · Latent variables · Human development

JEL Classification codes C3 · C43 · I32 · O15

1 Introduction

Many economic and social concepts such as welfare and poverty are multidimensional in nature and hence their operationalisation needs measures or indices that capture and combine the various dimensions in an adequate manner. Let us take the concept of human development for instance. This concept, first proposed by UNDP's Human Development Report in 1990 (see UNDP, HDR 1990) and largely inspired from Sen's various works (cf. e.g. Sen 1985, 1999), represents a major effort to reflect the multidimensional nature of well-being. The capability approach of Sen (re)defines development as the enhancement of

J. Krishnakumar (✉)
Department of Econometrics, University of Geneva, 40, Bd. du Pont d'Arve, 1211 Geneva 4, Switzerland
e-mail: jaya.krishnakumar@metri.unige.ch

A. L. Nagar
National Institute of Public Finance and Policy, New Delhi, India

people's choices or capabilities in various fields: economic, social, political, cultural and so on.

As it often happens for any new theoretical concept, the road from theory to practice is full of obstacles and we are still at the beginning of the road as far as capability approach is concerned. In fact there need not and will not be only one road but many paths leading to a practical measure of human development. In this case, a fundamental problem arises due to the fact that it is not possible to directly observe the concept as such. Almost all studies point to this feature and agree that there are many components to it. Hence it is often measured by means of several indicators and constructed as a composite index aggregating these indicators.

The earliest quantification attempts consisted in selecting different indicators and calculating a weighted average of these indicators. The most well-known among them are the Physical Quality of Life Index (PQLI) proposed by Morris (1979) and the above-mentioned Human Development Index (HDI) proposed by UNDP (1990). PQLI is a simple average of life expectancy at age one, infant mortality and adult literacy. HDI is similar but includes slightly different dimensions: health and longevity (measured by life expectancy at birth), instruction and access to knowledge (measured by literacy rate and enrollment ratio) and other dimensions for having a decent life (for which income is taken as a proxy). The three dimensions are given equal weights in the construction of the HDI. On the poverty side, we have the Human Poverty Index (UNDP, HDR 1997) which is a weighted average measuring deprivation in the same three dimensions of health (survival), education (illiteracy) and economic deprivation (itself a combination of three elements—access to health, safe water and adequate nourishment of children) for developing countries.

Two crucial issues in the above procedures are the adequacy of the chosen indicators for the corresponding dimension and the arbitrariness in the choice of weights. Over the recent years other indices have been proposed, derived from an underlying theoretical model, that offer an explanation for the inclusion of the variables composing the index as well as a better justification for the choice and values of the weights in the construction of the index.

These models are appealing because of two characteristics: (a) they assume that the underlying concept is not directly observable (i.e. is latent) but manifests itself in many observable quantities and (b) any single indicator can only be a partial measure of the underlying concept. Factor analysis, MIMIC (multiple indicators and multiple causes) and structural equation models (SEM) all fall into this line of reasoning. Latent variable models are common in psychology and the reader can find an excellent coverage of most of these models with applications in Bollen (1989), Bartholomew and Knott (1999), Muthen (2002) and Skrondal and Rabe-Hesketh (2004). Though principal components (PC) is not a latent variable model, it is widely used in empirical applications as an 'aggregating' technique and there is some confusion in the empirical literature which sometimes tends to equate principal components and latent factors. These two methods have different theoretical foundations and approach the problem from different angles. The principal components method is a pure data reduction technique that seeks linear combinations of the observed indicators in such a way as to reproduce the original variance as closely as possible. There is no underlying explanatory model in this method. On the other hand, the factor analysis is an explanatory model in which the observed values are postulated to be (linear) functions of a certain (fewer) number of unobserved latent variables (called factors). This paper examines the analytical expressions of the estimators derived from these two models (one descriptive and another explanatory) and shows that under certain special conditions the PC's are equivalent to the factor scores.

The MIMIC model (cf. Jöreskog and Goldberger 1975) represents a step further in the explanation of the phenomenon under investigation as it is not only believed that the observed variables are manifestations of a latent concept but also that there are other exogenous variables that "cause" and influence the latent factor(s). The structural equation model (SEM) goes beyond one-way causal links and specifies interdependencies among the latent variables while also including exogenous "causes". Thus it emphasises the simultaneous determination of the different (latent) dimensions of well-being while accounting for the impossibility of their direct measurement. We feel that it is the most suitable framework in the economic and social context as it provides one single index that incorporates in it the complex mechanisms involved in the formation of the latent concept that it is supposed to represent.

This paper reviews the most important latent variable models which form the basis of multidimensional indices of human development (or deprivation) starting from simpler ones such as factor analysis and going up to structural equation models. Only those features of each model that are relevant for our context, namely the construction of a multidimensional index, are presented in the review, directing the reader to related references for further details.

The next section presents the principal components method, the resulting index and its variance. It is followed by the factor analysis model in Sect. 3 where different possible indices (factor scores) are discussed and their properties derived. Section 4 derives the special conditions under which PC and FA can be seen to produce equivalent results. Thus it addresses the confusion in the empirical literature in the use of these two terms which should generally refer to two distinct quantities and clarifies the circumstances under which the terms can be used in an interchangeable manner. Section 5 examines the index and its properties in MIMIC models. Indices based on SEM are studied in the following section. Section 7 ends the chapter with a few concluding remarks.

2 Principal Components Indices

The use of principal components (PC) or a combination of principal components is a commonly used technique in the measurement of quality of life or well-being. This method, which is essentially a data reduction technique, dates back to Hotelling (1933) in the statistical literature with a wide range of applications in numerous fields such as psychology, biology, anthropology and more recently in economics and finance.

One of the earliest studies in the area of welfare is Ram (1982) who first applies PC on the three dimensions of PQLI mentioned above namely life expectancy at age one, infant mortality and adult literacy and combines it with per capita GDP, again using PC, to form a composite index. Slottje (1991) follows the same approach by selecting 20 attributes for 126 countries across the world, calculating a PC-based index and comparing it with indices obtained using hedonic weighting procedures. The PC method is still one of the most frequently used in empirical literature probably due to its computational simplicity (see e.g. Klasen 2000; Nagar and Basu 2001; Biswas and Caliendo 2002; Rahman et al. 2003; Noorbaksh 2003; McGillivray 2005).

The basic idea behind this method is to determine orthogonal linear combinations of a set of observed indicators chosen in such a way as to reproduce the original variance as closely as possible. Here we introduce some notations that will be used throughout the paper. Let y denote a $k \times 1$ vector of observed variables (which we already assume to be centered without loss of generality) and let Σ denote its covariance matrix. Let us further denote by $\theta_1, \ldots, \theta_k$ the k eigenvalues of Σ and by $a_1, \ldots, a_k$ the corresponding

eigenvectors. For the moment we assume Σ to be known (which will be replaced by its empirical version in practice). Then the principal components are given by:

$$p_j = a_j' y \quad j = 1, \ldots, k$$

or

$$p = A' y$$

where $A = [a_1 \ldots a_k]$ is the matrix of eigenvectors of Σ. We have $A'A = AA' = I_k$ and $\Sigma = A\Theta A'$ or $A'\Sigma A = \Theta$ where $\Theta = \text{diag}(\theta_j)$, $j = 1,\ldots k$ with the θ_j's arranged in descending order of magnitude. We also have $\Sigma^{-1} = A\Theta^{-1}A'$. The variances of the PC's are equal to the corresponding eigenvalues i.e. $V(p_j) = \theta_j \forall j$.

One of the interpretations that is often made regarding the principal components is that they are estimates of latent variables of which the observed values are indicators. It should be remembered that this method is originally a purely descriptive technique which tries to reproduce the observed variance or a large proportion of it using linear combinations. The above interpretation is in fact the underlying assumption for the factor analysis (FA) model to which we will turn in the next section.

Before going to the FA model and the link between PC and FA, let us present the indicators derived from PC's. The two most commonly used are the first principal component i.e. the one corresponding to the greatest eigenvalue θ_1 and a weighted average of all the principal components p_j's, $j = 1,\ldots,k$ with the weights w_j being given by the proportion of the total variance explained by each PC.

If we take the first principal component $p_1 = a_1' y$ as an aggregate index then we have $V(p_1) = \theta_1$. As for the weighted average its variance can be calculated as follows. Let us write it as:

$$\hat{H} = \sum_{j=1}^{k} w_j p_j$$

with

$$w_j = \frac{\theta_j}{\sum_{j=1}^{k} \theta_j}.$$

Denoting $\Theta = \text{diag}\,(\theta_j)$ and $w' = [w_1 \ldots w_k]$ and using $V(p_j) = \theta_j$ we have

$$V(\hat{H}) = w'\Theta w$$

where

$$w = (\iota'\Theta\iota)^{-1}\iota'\Theta.$$

Thus

$$V(\hat{H}) = \frac{\iota'\Theta^3\iota}{(\iota'\Theta\iota)^2}.$$

In practice, Σ is unknown and hence has to be estimated and the eigenvalues and eigenvectors of the estimator have to be used. These estimators are consistent (see e.g. Anderson 1984).

Though these indices are often used in empirical studies, few (none to our knowledge) give an estimation of their variance (or precision). Here we have a convenient expression that can be easily implemented.

3 Factor Analysis Model

The FA model assumes that the observed variables (indicators) are all dependent on one or more latent variables which are taken to be their common cause(s). Thus it not only conforms to our idea that the concept we are trying to assess is unobservable but also provides a theoretical framework explaining the observed variables as different manifestations of our latent concept(s) called factor(s). Some examples of works using factor analysis are Massoumi and Nickelsberg (1988), Schokkaert and Van Ootehgem (1990), Balestrino and Siclone (2000) and Lelli (2001).

The model is written as

$$y = \Lambda f + \varepsilon \tag{1}$$

where y $(k \times 1)$ denotes the vector of observed variables, f $(m \times 1)$ vector of latent variables $(m < k)$ and Λ the $(k \times m)$ coefficient matrix. If there is only one latent factor (for instance overall human development) then f is a scalar and Λ a $(k \times 1)$ vector.

Treating the latent factors as random, one assumes in general

$$V(f) = \Phi \quad \text{and} \quad V(\varepsilon) = \Psi$$

with Φ, Ψ positive definite. Let Σ denote the variance covariance matrix of the observed vector y as before. Then

$$\Sigma = \Lambda \Phi \Lambda' + \Psi.$$

This model uses the empirical estimators of Σ to find Λ, Φ and Ψ. It is usual to fix $\Phi = I$ for identification purposes. For the same reason, it is also assumed that $\Gamma = \Lambda' \Psi^{-1} \Lambda$ is diagonal. Maximum likelihood procedure is applied to the model to estimate Λ and Ψ given Σ. Given Λ, Ψ, one can derive minimum variance estimators or predictors of f as follows:

$$\hat{f} = (I + \Gamma)^{-1} \Lambda' \Psi^{-1} y \tag{2}$$

This estimator minimises $V(\hat{f} - f)$. It is also such that $\hat{f} = E(f|y)$ assuming joint normal distribution for (y,f).

Estimated in this way we do not have $E(\hat{f} - f|f)$ equal to zero. If we add it as a condition then we would obtain the following slightly different estimator (see Appendix A):

$$\hat{f}^* = \Gamma^{-1} \Lambda' \Psi^{-1} y = (\Lambda' \Psi^{-1} \Lambda)^{-1} \Lambda' \Psi^{-1} y \tag{3}$$

which is the least squares estimator of f in model (1) given y,Λ.

It can be argued that $E(\hat{f} - f|f) = 0$ may not be not a pertinent condition when f is not observed. In any case, the only difference between $\hat{f}$ and $\hat{f}^*$ is that $(I + \Gamma)$ in $\hat{f}$ is replaced by Γ in $\hat{f}^*$. Since Γ is diagonal this only means a rescaling of $\hat{f}$'s.

Let us now consider the special case $\Psi = I$. Then we get the following factor scores:

$$\tilde{f} = (I + \Lambda'\Lambda)^{-1}\Lambda' \tag{4}$$

and

$$\tilde{f}^{*} = (\Lambda'\Lambda)^{-1}\Lambda' y \tag{5}$$

for the 'unbiased' estimation.

4 Link between PC and FA Models

Let us take the case $\Psi = I$. Denoting the matrix of the m eigenvalues of Σ–I as Θ^{*} (note that $\theta_j^{*} = \theta_j - 1$ for $j = 1,...,m$) and using

$$A^{*}\Theta^{*}A^{*\prime} = \Sigma - I,$$

we can identify Λ as

$$\Lambda = A^{*}\Theta^{*\frac{1}{2}}$$

and write

$$\tilde{f} = (I + \Theta^{*})^{-1}\Theta^{*\frac{1}{2}}A^{*\prime}y = \Theta^{-1}\Theta^{*\frac{1}{2}}p^{*}$$

where p^{*} represents the first m principal components of Σ. Thus we see that the estimators of the latent variables obtained in the FA model are proportional to those given by the PC model (recalling that Θ^{*} and Θ are diagonal). For the 'unbiased' estimation, we have:

$$\tilde{f}^{*} = \Theta^{*-\frac{1}{2}}A^{*\prime}y = \Theta^{*-\frac{1}{2}}p^{*}.$$

Let us go a step further and consider the principal components to be potential estimators of the same latent factors as often done in empirical studies. Then requiring the PC's to be also 'unbiased' in the sense that $E(p^{**}\text{–}f|f) = 0$ yields (see Appendix B)

$$p^{**} = \Theta^{*\frac{1}{2}}A^{*\prime}y = \tilde{f}^{*}$$

The above identity between the 'unbiased' versions of PC's and factor scores not only completes the various links existing between PC and FA but also gives the theoretical justification behind the interpretation of principal components as latent variable estimators.

In case $\Psi \neq I$ but diagonal and one still wants to maintain the link between the two methods, then one has to premultiply the FA equation (1) by $\Psi^{-\frac{1}{2}}$ to obtain a new model

$$\Psi^{-\frac{1}{2}}y = \Psi^{-\frac{1}{2}}\Lambda f + \Psi^{-\frac{1}{2}}\varepsilon$$

or

$$\underline{y} = \underline{\Lambda} f + \underline{\varepsilon} \tag{6}$$

with a spherical $\underline{\varepsilon}$ and then apply PC or FA to the transformed model (6) for which the above result will hold. Note that the above transformation does not change the factor scores but only the factor loadings and needs a prior estimate of Ψ for its implementation. For this purpose one can use the ML estimate of Ψ obtained for the original (untransformed) FA model.

5 MIMIC Models

This model initially proposed by Jöreskog and Goldberger (1975) goes further in the theoretical explanation by introducing "causes" of latent factors. According to this model, the observed variables result from the latent factors and the latent factors themselves are caused by other exogenous variables denoted here as x. Thus we have a 'measurement equation' and a 'causal' relationship:

$$\begin{aligned} y &= \lambda f + \varepsilon \\ f &= \beta' x + \epsilon \end{aligned} \tag{7}$$

In their model with f a scalar and hence β, α, x vectors, the authors showed that the estimator of f is given by

$$\hat{f} = (1 - \lambda' \Omega^{-1} \lambda)^{-1} (\alpha' x + \lambda' \Psi^{-1} y)$$

with $V(\varepsilon) = \Psi$, $V(\varepsilon) = \sigma^2 I$, $\Omega = \lambda \lambda' + \Psi$.

The multivariate extension of this model is straightforward:

$$\begin{aligned} y &= \Lambda f + \varepsilon \\ f &= Bx + \epsilon \end{aligned} \tag{8}$$

with f a vector, Λ, B matrices of appropriate dimensions, and $V(\varepsilon) = \Psi$, $V(\varepsilon) = \sigma^2 I$.

Then we have

$$\hat{f} = (I - \Lambda' \Omega^{-1} \Lambda)^{-1} (Bx + \Lambda' \Psi^{-1} y).$$

Using the expression for the inverse of $\Omega = (\Psi + \Lambda\Lambda')$, one gets (see Appendix C)

$$\hat{f} = (I + \Lambda' \Psi^{-1} \Lambda)^{-1} Bx + (I + \Lambda' \Psi^{-1} \Lambda)^{-1} \Lambda' \Psi^{-1} y \tag{9}$$

The above equation shows that the MIMIC latent factor estimator is a sum of two terms: the first one is the "causes" term (function of x) and the second one can be called the "indicators" term. Note that the latter is nothing but the factor scores (2) of the FA model. If there are no 'causes' then (9) reduces to the pure FA estimator as one can expect.

Its variance is given by

$$V(\hat{f}) = BV(x)B' + (I + \Lambda' \Psi^{-1} \Lambda)^{-1} (\Lambda' \Psi^{-1} \Lambda)$$

Di Tommaso (2006) and Kuklys (2005) present two important applications of this methodology for welfare measurement. The former adopts the MIMIC approach to conceptualise children's well being using Indian data while the latter applies the MIMIC model for measuring the unobserved functioning in health and housing, each observed through a range of indicators and uses data from the British Household Panel Survey for 1991 and 2000 for estimating the model.

6 Structural Equation Models

Recall that the main idea behind the latent variable approach is that the different dimensions of development (or deprivation) cannot be directly measured but can be represented

by latent variables manifesting themselves through a set of achievements (or the lack of it). At the same time these latent dimensions mutually influence one another and hence it is important to explicitly specify these interactions in the form of a structural model.

Thus the most appropriate extension to the above models is an interdependent system of equations for the latent variables incorporating exogenous elements and a set of measurement equations linking the unobserved variables to the observed indicators. This is called the structural equation model (SEM), the most well-known in this category being the LISREL model proposed by Jöreskog (1973). This model specifies a system of equations explaining the latent variables (which become the endogenous variables of the model) by a set of exogenous (also latent) variables and including mutual effects of the endogenous variables on one another. To this system is added a set of equations to take account of the additional assumption that these latent endogenous and exogenous elements are observed through some indicators. This yields:

$$Ay^* + Bx^* + u = 0 \tag{10}$$

$$y = \Lambda y^* + \varepsilon \tag{11}$$

$$x = \Upsilon x^* + \epsilon \tag{12}$$

with

$$V(u) = \Sigma, \quad V(\varepsilon) = \Psi, \quad V(\epsilon) = \Xi$$

where (10) is the structural model and (11) and (12) constitute the measurement equations. We assume that the observations are centered without loss of generality.

Though (12) does not pose any additional problem on the theoretical side, we will remove it in the context of human development or well-being as the exogenous variables (basically representing institutional and social structures) will generally be observed. Though the statistical literature in this area has seen several extensions of the above model with ordinal/categorical variables and/or covariates (exogenous variables) in measurement equations (cf. Muthen 1984, 2002; Jöreskog 2002; Skrondal and Rabe-Hesketh 2004), we will continue with the above formulation for clarity of exposition.

The parameters of (10) and (11) can be estimated by generalised method of moments (GMM) by minimising the distance between the empirical variance covariance matrix of the y's and x's and the theoretical expressions of the covariance matrix given by (see e.g. Browne 1984):

$$V\left(\begin{bmatrix} y \\ x \end{bmatrix}\right) = \begin{bmatrix} (\Lambda A^{-1}(BV(x)B' + \Sigma)A'^{-1}\Lambda' + \Psi & \Lambda A^{-1}V(x) \\ V(x)A'^{-1}\Lambda' & V(x) \end{bmatrix}$$

and taking into account any a priori constraints on the parameters. The distance is optimally calculated in the metric (weight matrix) given by the inverse of the asymptotic variance–covariance matrix of the vector of sample statistics. This weighted least squares procedure is equivalent to a non-linear GMM procedure on the reduced form of the SEM.

An alternative procedure is the minimisation of the same distance between theoretical and empirical variance matrices conditioning on x. This is often the case as in general the mean and the variance of x are not restricted and are estimated by their sample values.

Then one would minimise the distance between the sample variance–covariance of y given x and ($\Lambda\, A^{-1}\Sigma\, A'^{-1}\, \Lambda' + \Psi$) under the same a priori constraints. Asymptotic theory gives us the variance matrix of the resulting estimators and a 'robust' version can be

computed to account for non-i.i.d. behaviour by estimating a heteroscedasticity-consistent estimate of the variance matrix.

One can also use (conditional) maximum likelihood (cf. e.g. Jöreskog 1973; Browne and Arminger 1995) to estimate the parameters under (conditional) normality of y^* given x and correct its variance using the well-known 'sandwich' formula under non-normality (quasi- maximum likelihood, cf. White 1982; Gourieroux et al. 1984).

Once the parameter estimates are obtained, the latent factors are estimated by their posterior means given the sample, replacing the parameter values by their estimates. This is called the Empirical Bayes estimator. For the above model (with observed x) we get (see Appendix D):

$$\hat{y}_i^* = A^{-1}Bx_i + A^{-1}\Sigma A^{-1'}\Lambda(\Lambda' A^{-1}\Sigma A^{-1'}\Lambda' + \Psi)^{-1}(y_i - \Lambda A^{-1}Bx_i)$$

or

$$\hat{y}_i^* = \left[I - A^{-1}\Sigma A^{-1'}\Lambda(\Lambda' A^{-1}\Sigma A^{-1'}\Lambda' + \Psi)^{-1}\Lambda\right]A^{-1}Bx_i + A^{-1}\Sigma A^{-1'}\Lambda(\Lambda' A^{-1}\Sigma A^{-1'}\Lambda' + \Psi)^{-1}y_i \tag{13}$$

From the point of view of a substantive interpretation of the above expression (13), it is important to point out that the factor scores are once again a combination of two terms: one capturing the 'causal' influence and the other reflecting the 'indicators' relevance.

Its variance can be obtained as (see Appendix D)

$$V(\hat{y}_i^*) = A^{-1}BV(x)B'A^{-1'} + A^{-1}\Sigma A^{-1'}\Lambda(\Lambda' A^{-1}\Sigma A^{-1'}\Lambda' + \Psi)^{-1}\Lambda A^{-1}\Sigma A^{-1}$$

An alternative method of obtaining factor scores is the maximum posterior likelihood which leads to the same result as (13) for our SEM given by (10), (11) (see Appendix D).

Note that the latent factors being ordinal, any monotonic increasing transformation of y^* will preserve the order in $\hat{y}^*$ (see Appendix E).

Let us end this section by citing a few major studies that apply the above model in the field of human development or poverty. Wagle (2005) uses a SEM for deriving multidimensional poverty measures using household data from a survey conducted in Kathmandu, Nepal in 2002 and 2003. Five major dimensions of well-being are considered: subjective economic well-being, objective economic well-being, economic well-being, economic inclusion, political inclusion and civic/cultural inclusion. Each of these dimensions is measured by a series of indicators and they influence one another through a system of simultaneous equations but there are no exogenous variables in the model. Krishnakumar (2007) proposes a general SEM with exogenous variables in both the structural and measurement parts for operationalising Sen's capability approach, including three dimensions namely knowledge, health and political freedom and demonstrates the utility of such a framework for deriving a multidimensional index of human development using worldwide country-level data. Krishnakumar and Ballon (2007) present another application of the same model using micro-level data on Bolivian households for analysing two basic capability domains—knowledge and living conditions.

7 Conclusions

It has become common to use multidimensional indices for measuring concepts such as well-being or poverty, in particular indices derived using principal components and latent

variable models. This paper brings out the motivation behind these approaches and their suitability in the economic and social domain. We begin with the PC method which is not a latent variable model but an entirely descriptive procedure for data reduction and hence useful for 'aggregating' several dimensions. The simplest latent variable model is the FA model which offers theoretical (measurement) relationships linking the observations and the latent dimensions. MIMIC structures add exogenous 'causes' for the latent variables. The SEM framework encompasses all these aspects and goes further in adding interdependencies and exogenous influences in both the structural and measurement equations.

Though the use of indices based on the above models and methods has become wide and popular, few studies report the standard errors or confidence intervals associated with these estimators. This paper examines their statistical properties, gives analytical expressions of their variances and establishes certain exact relationships among them.

Appendix A

Minimum Variance Unbiased Estimation of Factor Scores in the FA Model

We are interested in estimators of latent factors $\hat{f}$ such that

$$E(\hat{f} - f|f) = 0$$

and

$$V(\hat{f} - f) \text{ is minimal.}$$

Let us denote the estimator as $\hat{f} = Cy$. Then $E(\hat{f} - f) = E(C(\Lambda f + \varepsilon) - f) = (C\Lambda - I)E(f) = 0$ implies the following condition:

$$C\Lambda = I$$

Thus we need to solve the following program:
Minimise $V(\hat{f} - f) = (C\Lambda - I)(C\Lambda - I)' + C\Psi C'$ under the constraint

$$C\Lambda = I.$$

The Lagrangian is :

$$\begin{aligned} \pounds &= tr[C\Lambda - I)(C\Lambda - I)' + C\Psi C'] - \rho' \text{vec}(C\Lambda - I) \\ &= tr[C\Lambda - I)(C\Lambda - I)' + C\Psi C'] - \rho'(\Lambda' \otimes I)\text{vec}C - \rho' \text{vec}I \end{aligned}$$

Substituting the constraint in the objective function we get

$$\pounds = trC\Psi C' - \rho'(\Lambda \otimes I)\text{vec}C - \rho' \text{vec}I$$

The first order conditions are given by:

$$(\Psi' \otimes I)\text{vec}C - (\Lambda \otimes I)\rho = 0$$

$$(\lambda' \otimes I)\text{vec}C = 0$$

Solving the above system, one obtains:

$$\rho^* = (\Lambda'\Psi^{-1}\Lambda)^{-1}$$
$$C^* = (\Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'\Psi^{-1}$$

In the special case $\Psi = I$, $C^* = (\Lambda'\Lambda)^{-1}\Lambda'$ and $\tilde{f} = C^*x = \Theta^{-\frac{1}{2}}A'x = \Theta^{-\frac{1}{2}}p$.

Appendix B

"Unbiased" Principal Components

If we require the first m principal components to be also unbiased estimators of the latent factors that they are supposed to represent then we should find B such that

$$E(BA^{*'}y - f|f) = 0 \quad \text{i.e. } E((BA^{*'}\Lambda - I)f|f) = 0 \quad \forall f.$$

This implies

$$BA^{*'}\Lambda - I = 0$$

or

$$BA^{*'}A^*\Theta^{*\frac{1}{2}} = I$$

or

$$B\Theta^{*\frac{1}{2}} = I$$

or

$$B = \Theta^{*-\frac{1}{2}}$$

Thus the 'unbiased' principal component estimator is given by

$$p^{**} = \Theta^{*-\frac{1}{2}}A'x = \Theta^{*-\frac{1}{2}}p = \tilde{f}^*.$$

Appendix C

Expression of MIMIC Estimator

Following Jöreskog and Goldberger (1975), the conditional expectation of f given y,x is given by:

$$\hat{f} = Bx + \Lambda'\Omega^{-1}(y - \Lambda Bx)$$

where

$$\Omega = \Lambda\Lambda' + \Psi$$

Using

$$(\Lambda\Lambda' + \Psi)^{-1} = \Psi^{-1} + \Psi^{-1}\Lambda(I + \Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'\Psi^{-1}$$

we obtain

$$\hat{f} = [I - \Lambda'\Psi^{-1}\Lambda + \Lambda'\Psi^{-1}\Lambda(I + \Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'\Psi^{-1}\Lambda][Bx + \Lambda'\Psi^{-1}y]$$

which can be simplified to

$$(I + \Lambda'\Psi^{-1}\Lambda)^{-1}(Bx + \Lambda'\Psi^{-1}y)$$

Appendix D

Latent Factor Estimators and Their Variances in the Linear SEM

As explained in the text, the latent factors are estimated as the expectation of the posterior distribution of these factors given the sample i.e. given y,x. For a pure measurement model (with exogenous variables w) written as

$$\begin{aligned} y &= Dw + \Lambda\eta + \varepsilon \\ x &= \eta_x \end{aligned} \tag{14}$$

the latent factor (Empirical Bayes) estimator is derived in Skrondal and Rabe-Hesketh (2004) as follows:

$$\hat{\eta} = V(\eta)\Lambda'(\Lambda V(\eta)\Lambda' + \Psi)^{-1}(y - Dw)$$

Here we take the above formula and adapt it to our case in which we have a SEM for explaining the latent factors. Our model is reproduced below for reference:

$$\begin{aligned} Ay^* + Bx^* + u &= 0 \\ y &= \Lambda y^* + \varepsilon \end{aligned} \tag{15}$$

with

$$V(u) = \Sigma$$

To make use of the above result we substitute the reduced form of our SEM given by

$$y^* = A^{-1}Bx + A^{-1}u$$

into the measurement equation (15) to get

$$y = \Lambda A^{-1}Bx + \Lambda A^{-1}u + \varepsilon \tag{16}$$

Identifying (16) with (14) and η with u one can obtain the 'estimator' of u as

$$\hat{u} = \Sigma A^{-1}\Lambda'(\Lambda A^{-1}\Sigma A^{-1'}\Lambda' + \Psi)^{-1}(y - \Lambda A^{-1}Bx)$$

The factor estimators are then obtained by substituting $\hat{u}$ for u in the SEM model (15):

$$\hat{y}^* = A^{-1}Bx + A^{-1}\Sigma A^{-1}\Lambda'(\Lambda A^{-1}\Sigma A^{-1'}\Lambda' + \Psi)^{-1}(y - \Lambda A^{-1}Bx) \tag{17}$$

which is the equation given in the text.

Finally, the variance of $\hat{y}^*$ is derived by noting that

$$y - \Lambda A^{-1}Bx = \Lambda A^{-1}u + \varepsilon$$

and

$$V(\Lambda A^{-1}u + \varepsilon) = \Lambda A^{-1}\Sigma A^{-1'}\Lambda' + \Psi$$

and using the above to calculate $V(\hat{y}^*)$ according to (17).

Alternatively, Muthen (1998-2004) gives another expression of the latent factor estimator based on maximisation of posterior likelihood. The model is written as

$$v = v_v + \Lambda_v \eta_v + \varepsilon_v$$

$$A_v \eta_v = \alpha_v + u_v$$

where

$$v = \begin{bmatrix} y \\ x \end{bmatrix}; \quad v_v = \begin{bmatrix} v_y \\ 0 \end{bmatrix} \quad \Lambda_v = \begin{bmatrix} \Lambda & 0 \\ 0 & I \end{bmatrix}; \quad \eta_v = \begin{bmatrix} \eta \\ \eta_x \end{bmatrix} \quad \varepsilon_v = \begin{bmatrix} \varepsilon \\ 0 \end{bmatrix}$$

$$A_v = \begin{bmatrix} A & -B \\ 0 & I \end{bmatrix}; \quad \alpha_v = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}; \quad u_v = \begin{bmatrix} u \\ 0 \end{bmatrix};$$

with

$$E(\varepsilon) = 0 \qquad E(u) = 0$$

and

$$V(\varepsilon) = \Psi \qquad V(u) = \Sigma$$

Thus the model is in fact

$$y = v + \Lambda\eta + \varepsilon$$
$$A\eta = \alpha + Bx + u$$

$$x = \eta_x$$

The factor score estimator is then:

$$\hat{\eta_v} = \mu_v + C(v - v_v - \Lambda_v \mu_v) \tag{18}$$

where

$$\mu_v = A^{-1}\alpha_v$$

$$C = A_v^{-1}\Sigma_v A_v^{-1'}\Lambda_v'(\Lambda_v A_v^{-1}\Sigma_v A_v^{-1'}\Lambda_v' + \Psi_v)^{-1}$$

and

$$\Sigma_v = \begin{bmatrix} \Sigma & 0 \\ 0 & \Sigma_{xx} \end{bmatrix}; \quad \Psi_v = \begin{bmatrix} \Psi & 0 \\ 0 & 0 \end{bmatrix}.$$

Replacing the above partitioned matrices and vectors in (18) and performing all the calculations, one gets:

$$\hat{\eta} = A^{-1}\alpha + A^{-1}Bx + A^{-}\Sigma A^{-1'}\Lambda(\Lambda A^{-1}\Sigma A^{-1'}\Lambda' + \Psi)^{-1}(y - v_y - \Lambda A^{-1}\alpha - \Lambda Bx)$$

and

$$\hat{\eta}_x = x$$

The last result is expected as we assume that the x's are directly observed.

Assuming y is centered and regrouping the intercept term $A^{-1}\alpha$ and the 'exogenous' elements term $A^{-1}Bx$ into one term denoting it with the same symbol $A^{-1}Bx$ (i.e. assuming x incorporates a constant), one gets

$$\hat{\eta} = A^{-1}Bx + A^{-1}\Sigma A^{-1'}\Lambda(\Lambda A^{-1}\Sigma A^{-1'}\Lambda' + \Psi)^{-1}(y - \Lambda A^{-1}Bx)$$

Thus we see that it is the same expression as the Empirical Bayes estimator (17) (under our above assumptions) and hence has the same variance.

Appendix E

Monotonic Transformation and Posterior Distribution

The ordinality of latent factors implies that any monotonic transformation of y^* will preserve the order in $\hat{y^*}$. We will show this in the case of a scalar latent factor y^* with a vector indicator y. The proof can be extended to the vector case without any major difficulty.

The posterior distribution of the latent factor y^* given the indicator y is given by

$$p(y^*|y) = \frac{p(y^*)\pi(y|y^*)}{f(y)}$$

where $p(y^*|y)$ denotes the posterior density of y^* given y, $p(y^*)$ is the prior density of y^*, $\pi(y|y^*)$ is the distribution of y given y^* and $f(y)$ denotes the density of y.

Let us now transform y^*: $u^* = g(y^*)$.

Then, using

$$y^* = g^{-1}(u^*), \quad p(u^*) = p(y^*)\left(\frac{dg}{dy^*}\right)^{-1}$$

and

$$\pi(y|y^*) = \pi(y|g^{-1}(u^*))$$

one can write

$$p(y^*|y) = \frac{p(y^*)\left(\frac{dg}{dy^*}\right)\pi(y|g^{-1}(u^*))}{f(y)}$$

or

$$= \left(\frac{dg}{dy^*}\right)\frac{p(g^{-1}(u^*))\left(\frac{dg}{dy^*}\right)\pi(y|g^{-1}(u^*))}{f(y)}$$

The first element of the product is positive if $g(y^*)$ is monotonic increasing and one can write the second part as $p(g^{-1}(u^*)|y) \equiv p(u^*|y)$.

Hence

$$p(u^*|y) = \left(\frac{dg}{dy^*}\right)^{-1} p(y^*|y)$$

Therefore if

$$E(y^*|y_1) > E(y^*|y_2)$$

then we have

$$\int y^* p(y^*|y_1)dy^* > \int y^* p(y^*|y_2)dy^*$$

$$\int g(y^*)p(y^*|y_1)dy^* > \int g(y^*)p(y^*|y_2)dy^*$$

$$\int g(y^*)\left(\frac{dg}{dy^*}\right)^{-1} p(u^*|y_1)dy^* > \int g(y^*)\left(\frac{dg}{dy^*}\right)^{-1} p(u^*|y_2)dy^*$$

$$\int u^*\left(\frac{dg}{dy^*}\right)^{-1} p(u^*|y_1)\left(\frac{dg}{dy^*}\right)du^* > \int u^*\left(\frac{dg}{dy^*}\right)^{-1} p(u^*|y_2)\left(\frac{dg}{dy^*}\right)du^*$$

$$\int u^* p(u^*|y_1)du^* > \int u^* p(u^*|y_2)du^*$$

and finally

$$E(u^*|y_1) > E(u^*|y_2).$$

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis*. New York: John Wiley and Sons.

Balestrino, A., & Sciclone, N. (2000). Should we use functionings instead of income to measure well-being? Theory, and some evidence from Italy. *Mimeo*, University of Pisa.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis*. U.K.: Edward Arnold.

Biswas, B., & Caliendo, F. (2002). A multivariate analysis of the human development index. *Indian Economic Journal, 49*(4), 96–100.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Staistical Psychology, 37*, 62–83.

Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean - and covariance-structural models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modelling for the social and behavioral sciences* (pp. 311–359). Newbury Park: Plenum Press.

Di Tommaso, M. L. (2006). Measuring the well-being of children using a capability approach: An application to Indian data. Working Paper CHILD No. 05/2006, University of Turin.

Gouriroux, C., Monfort, A., & Trognon, A.(1984). Pseudo-maximum likelihood methods: Theory. *Econometrica, 52*, 681–700.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*, 417–441

Jöreskog, K. (1973). A general method for estimatimg a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences*. New York: Seminar Press.

Jöreskog, K. (2002). Structural equation modelling with ordinal variables using LISREL. http://www.ssicentral.com/lisrel/ordinal.htm

Jöreskog, K., & Goldberger, A. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70*(351), 631–639.

Klasen, S. (2000). Measuring poverty and deprivation in South Africa. *Review of Income and Wealth, 46*, 33–58.

Krishnakumar, J. (2007). Going beyond functionings to capabilities: An econometric model to explain and estimate capabilities. *Journal of Human Development, 7*, 39–63.

Krishnakumar, J., & Ballon, P. (2007). Estimating basic capabilities: A structural equation model applied to Bolivia. Working paper under review.

Kuklys, W. (2005). *Amartya Sen's capability approach: Theoretical insights and empirical applications*. Berlin: Springer.

Lelli, S. (2001). Factor analysis vs. fuzzy sets theory: Assessing the influence of different techniques on Sen's functioning approach. Center for Economic studies, K.U. Leuven.

Maasoumi, E., & Nickelsburg, G. (1988). Multidimensional measures of well-being and an analysis of inequality in the Michigan data. *Journal of Business and Economic Statistics, 6*(3), 327–334.

McGillivray, M. (2005). Measuring non-economic well-being achievement. *Review of Income and Wealth, 51*(2), 337–364.

Morris, M. D. (1979). *Measuring the condition of the world's poor: The physical quality of life index*. New York: Pergamon.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent indicators. *Psychometrika, 49*, 115–132.

Muthen, B. (2002). Beyond SEM: General latent variable modelling. *Behaviormetrika, 29*(1), 81–117.

Muthen, B. O. (1998-2004). *Mplus technical appendices*. Los Angeles, CA: Muthen & Muthen.

Nagar, A. L., & Basu, S. (2001). *Weighting socio-economic indicators of human development (a latent variable approach)*. New Delhi: National Institute of Public Finance and Policy.

Noorbaksh, F. (2003). Human development and regional disparities in India. Discussion Paper, Helsinki: UN-WIDER.

Rahman, T., Mittelhammer, R. C., & Wandschneider, P. (2003). Measuring the quality of life across countries: A sensitivity analysis of well-being indices. In *WIDER International Conference on Inequality, Poverty and Human Well-being*, Helsinki, Finland

Ram, R. (1982). Composite indices of physical quality of life, basic needs fulfilment, and income: A principal component representation. *Journal of Development Economics, 11*, 227–247.

Schokkaert, E., & Lootehgem, L. (1990). Sen's concept of the living standard applied to the Belgian unemployed. *Recherches Economiques de Louvain, 56*, 429–450.

Sen, A. K. (1985). *Commodities and capabilities*. Amsterdam: North-Holland.

Sen, A. K. (1999). *Development as freedom*. Oxford: Oxford University Press.

Slottje, D. J (1991). Measuring the quality of life across countries. *The Review of Economics and Statistics 73*(4), 684–693.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, U.S.A.: Chapman & Hall/CRC.

UNDP (1990). *Human Development Report (HDR)*. U.K.: Oxford University Press.

Wagle, U. (2005). Multidimensional poverty measurement with economic well-being, capability and social inclusion: A case from Kathmandu, Nepal. *Journal of Human Development, 6*(3), 301–328.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*, 1–26.